COMPUTATIONAL METHODS IN APPLIED MATHEMATICS, Vol.3(2003), No.3, pp.493–512

ON CONVERGENCE OF THE EXPONENTIALLY FITTED FINITE VOLUME METHOD WITH AN ANISOTROPIC MESH REFINEMENT FOR A SINGULARLY PERTURBED CONVECTION-DIFFUSION EQUATION

SONG WANG

School of Mathematics & Statistics, The University of Western Australia
35 Stirling Highway, Crawley, WA 6009, Australia

LUTZ ANGERMANN

Institut für Mathematik, Technische Universität Clausthal
Erzstraße 1, D-38678 Clausthal-Zellerfeld, Germany

Dedicated to John J.H.Miller on the occasion of his 65th birthday.

Abstract — This paper presents a convergence analysis for the exponentially fitted finite volume method in two dimensions applied to a linear singularly perturbed convection-diffusion equation with exponential boundary layers. The method is formulated as a nonconforming Petrov-Galerkin finite element method with an exponentially fitted trial space and a piecewise constant test space. The corresponding bilinear form is proved to be coercive with respect to a discrete energy norm. It is also shown that the approximation error in the discrete energy norm is bounded above by $C\left(h^{1/2}+h\sqrt{|\ln\varepsilon/\ln h|}\right)$ with C independent of the mesh parameter h, the diffusion coefficient ε, and the exact solution of the problem. Numerical results are presented to verify the theoretical rates of convergence.

2000 Mathematics Subject Classification: 65N30, 76M10.

Keywords: finite volume method, convection-diffusion equation, uniform convergence, anisotropic mesh, singular perturbation.

Dedication: The authors would like to dedicate this work to John Miller for his many pioneer contributions to numerical solution of singular perturbation problems and semiconductor device equations. The present work is a continuation of several papers co-authored by him.

1. Introduction

The exponentially fitted finite volume method, known as the Scharfetter-Gummel box integration method, is the most popular method for solving the partial differential equations in

the drift-diffusion model of semiconductor devices, to which the solutions display sharp layers [7,9,10,14]. This method is based on the idea proposed by Scharfetter and Gummel [19] to approximate a flux in a small interval by a constant, leading to a local exponential approximation to the potential function. This is in contrast to a conventional finite element method in which the potential function is approximated locally by, say, a linear function, yielding a constant approximation to the gradient of the potential. The same idea was also proposed in [1] for solving a fluid flow problem. An intuitive reason that the Scharfetter-Gummel technique works well is that, in practice, a flux behaves better than the gradient of the corresponding potential function. The one-dimensional Scharfetter-Gummel's method has been extended to higher dimensions by many researchers (cf., for example, [3,9,23]) and some of these extensions have been used for solving singularly perturbed and incompressible Navier-Stokes equations (cf., for example, [2,12,13,22]). Although the method has been used successfully for solving a number of problems, very little theoretical work on the stability and convergence analysis of the method on unstructured meshes can be found in the open literature. The first analysis of this method was given in [15]. More refined mathematical analysis for problems with boundary and interior layers can be found in [12] and [14]. But the upper error bounds established in these works depend strongly and unfavorably on the singular perturbation parameters. A uniform convergence analysis for the case of uniform rectangular meshes is given in [11]. On the other hand, there are many notable advances on the uniform convergence analysis of other methods such as the streamline diffusion method and the standard piecewise bilinear finite element method (cf., for example, [20, 21, 25]), although most of these works are based on structured piecewise uniform rectangular meshes. Therefore, unlike other methods, the mathematical understanding of the exponentially fitted finite volume method is very limited.

In the present paper, we study in detail the stability and convergence properties of the method on an unstructured 2D mesh with an anisotropic refinement when the method is applied to a linear singularly perturbed convection-diffusion problem with a singular perturbation parameter ε. We show that the method is numerically stable in the sense that the corresponding bilinear form is coercive with respect to a discrete energy norm which is virtually independent of ε. We also show that the error in the discrete norm is bounded above by $\mathcal{O}(h^{1/2})$. This error bound is almost independent of ε and provides a uniform convergence of order $h^{1/2}$ for the range of all practical values of ε. The rest of the paper is organized as follows.

The continuous problem and some a priori estimates on the exact solution and its derivatives are discussed in the next section. In Section 3, we formulate the finite volume method as a Petrov-Galerkin finite element method and transform the Petrov-Galerkin method into a Bubnov-Galerkin one. We will show in Section 4 that the method is numerically stable by demonstrating that the bilinear form is coercive with respect to a discrete energy norm. We will also present an error analysis for the finite element solution and show that the global error of the approximation in the discrete energy norm is bounded above by $\mathcal{O}(h^{1/2})$ almost uniformly in ε. In Section 5 we will present some numerical results to verify the theoretical rates of convergence. The numerical results also demonstrate the superconvergence phenomenon of the method when a piecewise uniform mesh is used, though it is not theoretically proved in this paper.

Although the analysis is performed in two dimensions, the idea is applicable to the three-dimensional case, too.

2. Preliminaries

The problem we consider in this paper is the following stationary, linear, convection-diffusion problem:

$$-\nabla \cdot \boldsymbol{f} + Gu = F \quad \text{in} \quad \Omega := (0,1)^2, \tag{2.1}$$

$$\boldsymbol{f} = \varepsilon \nabla u - \boldsymbol{a}u, \tag{2.2}$$

$$u|_{\partial\Omega} = 0, \tag{2.3}$$

where $\partial\Omega$ denotes the boundary of Ω, $\varepsilon > 0$ is a positive parameter, $\boldsymbol{a} = (a_1, a_2)$ is a known vector-valued function, and F and G are given function.

Before making assumptions on the given functions, we first introduce some notation of function spaces. In what follows $L^p(S)$ denotes the space of p-integrable functions on an open and measurable set S with the norm $\|\cdot\|_{0,p,S}$ and $W^{m,p}(\Omega)$ the usual Sobolev space with the norm $\|\cdot\|_{m,p,S}$ and the kth order seminorm $|\cdot|_{k,p,S}$ for any $1 \leqslant p \leqslant \infty$, the nonnegative integers m and k satisfying $0 \leqslant k \leqslant m$. Obviously, $W^{0,p}(S) = L^p(S)$. When $S = \Omega$, we omit the subscript in the above notation. Furthermore, we let $H^m(\Omega) := W^{m,2}(\Omega)$, $||\cdot||_m := ||\cdot||_{m,2,\Omega}$ and $|\cdot|_k := |\cdot|_{k,2,\Omega}$. The inner product on $L^2(\Omega)$ or on $\boldsymbol{L}^2(\Omega) := (L^2(\Omega))^2$ is denoted by $(\cdot,\cdot)$. We put $H_0^1(\Omega) := \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}$ and the set of functions which together with their up to and including m order derivatives are continuous on Ω (or $\bar{\Omega}$) is denoted by $C^m(\Omega)$ (or $C^m(\bar{\Omega})$). We use $|\cdot|$ to denote an absolute value, Euclidean length, or area depending on the context.

For the coefficient functions we assume that $\boldsymbol{a} \in (C^1(\Omega))^2$, $G \in C^0(\overline{\Omega}) \cap H^1(\Omega)$ and $F \in L^2(\Omega)$. We also assume that $\boldsymbol{a}$ and G satisfy

$$\frac{1}{2}\nabla \cdot \boldsymbol{a} + G \geqslant 0 \qquad \text{in} \quad \Omega. \tag{2.4}$$

Furthermore, we assume that the components of $\boldsymbol{a}$ are bounded below by two positive constants α_1 and α_2, respectively, i.e.,

$$a_1 \geqslant \alpha_1 > 0, \qquad a_2 \geqslant \alpha_2 > 0 \qquad \text{in} \quad \Omega. \tag{2.5}$$

We also assume that $\varepsilon \ll |\boldsymbol{a}|$ so that, in this case, the solution to (2.1) – (2.3) has two exponential boundary layers of width $\mathcal{O}(\varepsilon)$ at $x = 1$ and $y = 1$. The variational problem corresponding to (2.1), (2.2) and (2.3) is

Problem 2.1. *Find $u \in H_0^1(\Omega)$ such that for all $v \in H_0^1(\Omega)$*

$$A(u,v) = (F,v),$$

where $A(\cdot,\cdot)$ is a bilinear form on $(H_0^1(\Omega))^2$ defined by

$$A(u,v) := (\varepsilon\nabla u - \boldsymbol{a}u, \nabla v) + (Gu, v).$$

Let $||\cdot||_\varepsilon$ be a functional on $H_0^1(\Omega)$ defined by $||v||_\varepsilon := (A(v,v))^{1/2}$. Then, it is easy to show that (cf., for example, [12])

$$||v||_\varepsilon^2 = (\varepsilon\nabla v, \nabla v) + \left(\left(\frac{1}{2}\nabla \cdot \boldsymbol{a} + G\right)v, v\right).$$

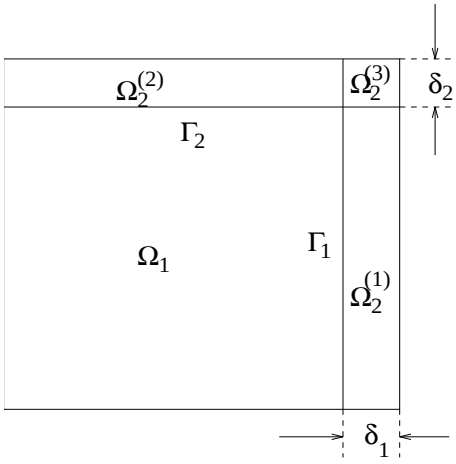

Figure 1. Subdomains Ω_1 and $\Omega_2 = \Omega_2^{(1)} \cup \Omega_2^{(2)} \cup \Omega_2^{(3)}$, $\Gamma = \Gamma_1 \cup \Gamma_2$

Thus, $||\cdot||_\varepsilon$ is a norm on $H_0^1(\Omega)$ because of the fact that $\frac{1}{2}\nabla \cdot \boldsymbol{a} + G \geqslant 0$ by (2.4) and that $(\nabla u, \nabla v)$ is a norm on $H_0^1(\Omega)$ by the well-known Poincaré-Friedrichs inequality. Now, from the definition of the norm we have

$$A(u,u) = ||u||_\varepsilon^2, \qquad \forall u \in H_0^1(\Omega).$$

This implies that $A(\cdot,\cdot)$ is coercive on $H_0^1(\Omega)$ and thus, by the well-known Lax-Milgram Lemma, Problem 2.1 has a unique solution in $H_0^1(\Omega)$.

Because of (2.5), the solution to Problem 2.1 has two boundary layers of width $\mathcal{O}(\varepsilon)$ at $x = 1$ and $y = 1$, respectively. Thus, we divide the solution region Ω into two parts Ω_1 and Ω_2 given respectively by

$$\Omega_1 = (0, 1-\delta_1) \times (0, 1-\delta_2) \quad \text{and} \quad \Omega_2 = (1-\delta_1, 1) \times (0,1) \cup (0, 1-\delta_1) \times (1-\delta_2, 1),$$

with $\delta_1, \delta_2 \in (0,1)$ (cf. Fig. 1). Obviously $\bar{\Omega}_1 \cup \bar{\Omega}_2 = \bar{\Omega}$. The region Ω_2 consists of three subregions $\Omega_2^{(1)}, \Omega_2^{(2)}$, and $\Omega_2^{(3)}$ defined respectively by

$$\begin{aligned}
\Omega_2^{(1)} &= (1-\delta_1, 1) \times (0, 1-\delta_2),\\
\Omega_2^{(2)} &= (0, 1-\delta_1) \times (1-\delta_2, 1),\\
\Omega_2^{(3)} &= (1-\delta_1, 1) \times (1-\delta_2, 1).
\end{aligned}$$

The choice of the transition parameters δ_1 and δ_2 is rather arbitrary, but it is required that Ω_2 should cover the boundary layers and $\delta_1, \delta_2 = O(\varepsilon)$. One choice is

$$\delta_1 = \frac{\beta}{\alpha_1}\varepsilon \ln\frac{1}{\varepsilon} \quad \text{and} \quad \delta_2 = \frac{\beta}{\alpha_2}\varepsilon \ln\frac{1}{\varepsilon}, \tag{2.6}$$

where $\beta \geqslant 1$ is a positive parameter (cf., for example, [18]). We let $\Gamma = \bar{\Omega}_1 \cap \bar{\Omega}_2$ which contains two segments Γ_1 and $\Gamma 2$.

Let us now consider the properties of the solution to Problem 2.1. Obviously, it is smooth in Ω_1 and has large derivatives along the x axis, y axis or both respectively in $\Omega_2^{(1)}$, $\Omega_2^{(2)}$, and $\Omega_2^{(3)}$. More precisely, the solution can be decomposed into four parts U_i $(i = 1,2,3,4)$ as given in the following assumption:

Assumption 2.1. *The solution u to Problem 2.1 has the representation*

$$u = U_1 + U_2 + U_3 + U_4,$$

where U_1 satisfies

$$||U_1||_{k,\infty,\Omega} \leqslant C \qquad for \quad k = 0, 1, 2, \tag{2.7}$$

and U_2, U_3, and U_4 satisfy

$$\left|\frac{\partial^{i+j} U_2}{\partial x^i \partial y^j}\right| \leqslant C\varepsilon^{-i} \exp\left(-\frac{\alpha_1(1-x)}{\varepsilon}\right), \tag{2.8}$$

$$\left|\frac{\partial^{i+j} U_3}{\partial x^i \partial y^j}\right| \leqslant C\varepsilon^{-j} \exp\left(-\frac{\alpha_2(1-y)}{\varepsilon}\right), \tag{2.9}$$

$$\left|\frac{\partial^{i+j} U_4}{\partial x^i \partial y^j}\right| \leqslant C\varepsilon^{-(i+j)} \exp\left(-\frac{\alpha_1(1-x)}{\varepsilon}\right) \exp\left(-\frac{\alpha_2(1-y)}{\varepsilon}\right), \tag{2.10}$$

for $0 \leqslant i + j \leqslant 2$ and some positive constant C.

The part U_1 is globally smooth and uniformly bounded, while U_2, U_3, and U_4 contain layers in $\Omega_2^{(1)}, \Omega_2^{(2)}$, and $\Omega_2^{(3)}$ respectively. Sufficient conditions for the existence of this decomposition have been discussed in various papers and books such as [8, 16], but necessary and sufficient conditions are unknown. The following theorem shows that u and all its first and second partial derivatives are uniformly bounded in Ω_1.

Theorem 2.1. *If $\beta \geqslant 2$, then*

$$||u||_{i,\infty,\Omega_1} \leqslant C \qquad i = 0, 1, 2.$$

Proof. The proof of this theorem follows directly from (2.6) and (2.7) – (2.10). □

3. The finite element formulation of the box method

For the formulation of the exponentially fitted finite box method as a finite volume method, we refer to [9, 13]. In this section we reformulate it as a Petrov-Galerkin finite element method, and then as an equivalent Bubnov-Galerkin method. This allows us to investigate the stability and error bounds of the box method using the framework of a finite element method.

To formulate the finite volume method as a finite element one, we first define a mesh for the solution region Ω which is a combination of triangles in Ω_1 and rectangles in Ω_2. Let $T_h^{(1)}$ denote a triangular mesh on Ω_1 with each triangle t having the diameter h_t less than or equal to h. We denote the sets of vertices and edges in $T_h^{(1)}$ by $X_h^{(1)}$ and $E_h^{(1)}$, respectively. For all $0 < h < \dim(\Omega_1)$, $\{T_h^{(1)}\}$ forms a family of triangular meshes on Ω_1. For each $T_h^{(1)}$ in the family we assume that $T_h^{(1)}$ is quasiuniform, *i.e.*, there exists a constant $\gamma > 0$ independent of h such that

$$\frac{\min_{e \in E_h^{(1)}} |e|}{h} \geqslant \gamma.$$

We use N_1 to denote the number of vertices of $T_h^{(1)}$ not on $\partial\Omega$.

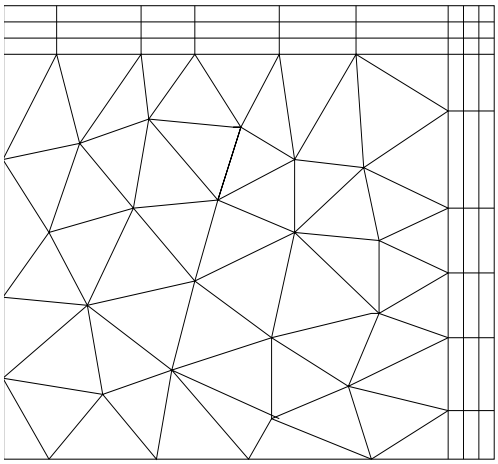

Figure 2. A typical hybrid mesh for Ω

The subregion Ω_2 contains two thin overlapped stripes $\Omega_2^{(1)} \cup \Omega_2^{(3)}$ and $\Omega_2^{(2)} \cup \Omega_2^{(3)}$ with the widths δ_1 and δ_2, respectively. Thus, we divide these two strips into rectangles so that the resulting mesh is uniform along the x axis with M_1 subintervals in the former subregion and along the y axis with M_2 subintervals in the latter (cf., Fig 2). We also require that the mesh points on Γ should match those from $T_h^{(1)}$. This mesh is denoted by $T_h^{(2)}$. Without loss of generality, we assume that the vertices in $T_h^{(2)}$ not on $\partial\Omega$ are numbered from $N_1 + 1$ to $N_1 + N_2$. The set of edges of $T_h^{(2)}$ not on $\partial\Omega$ is denoted by $E_h^{(2)}$. Obviously all rectangles in $\Omega_2^{(1)}$ and $\Omega_2^{(2)}$ have lengths $\mathcal{O}(h)$ and widths either δ_1/M_1 or δ_2/M_2, and rectangles in $\Omega_2^{(3)}$ have the length δ_1/M_1 and width δ_2/M_2. The meshes $T_h^{(1)}$ and $T_h^{(2)}$ form a conforming mesh on Ω and we denote it by T_h. A typical case is depicted in Fig. 2. We let the total number of nodes of T_h not on $\partial\Omega$ be N' which equals $N_1 + N_2$ minus the number of nodes on Γ.

We comment that this mesh is not a Shishkin type of mesh, because the transition width along each direction is $\mathcal{O}(\varepsilon \ln(1/\varepsilon))$ rather than $\mathcal{O}(\varepsilon \ln \bar{N})$, where $\bar{N}$ denotes the number of mesh points in the direction. Since in this paper we are interested in error analysis of the finite volume method on unstructured meshes, it is not possible to define $\bar{N}$ exactly. As can be seen later, the main error upper error bound obtained in this paper still depends (very weakly) on ε.

Now, we let X_h and E_h denote respectively the sets of vertices and edges of T_h and let X_h' and E_h' denote respectively the sets of vertices and edges of T_h not on $\partial\Omega$.

Definition 3.1. *T_h is a Delaunay mesh if, for every element $t \in T_h$, the circumcircle of t contains no other vertices in X_h (cf., [5]).*

We assume henceforth that each T_h contains a Delaunay mesh. This allows us to construct two meshes dual to T_h as given below.

Associated with T_h, we define two meshes dual to it. The first dual mesh, denoted by D_h, is the Dirichlet tessellation (cf., [6]) associated with the mesh nodes in T_h, i.e. the element $d_i \in D_h$ associated with the node x_i of T_h is given by

$$d_i = \{x \in \Omega : |x - x_i| \leqslant |x - x_j|, i \neq j\}$$

for any other node x_j of the mesh T_h. The assumption that T_h is a Delaunay mesh guarantees that the Dirichlet tessellation D_h dual to T_h exists.

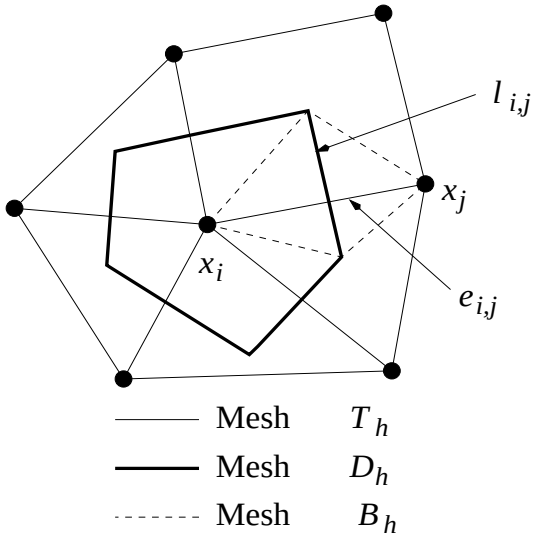

Figure 3. Elements and edges associated with the node x_i

For each edge in E'_h, we construct a quadrilateral element by connecting the two endpoints of the edge and the circumcenters of the two elements sharing the edge. All these quadrilaterals form the second dual mesh denoted by B_h. An example of these nested meshes is depicted in Fig. 3.

Using the meshes defined above, we now construct finite element trial and test spaces, $U_h, V_h \subset L^2(\Omega)$, each of dimension N', the number of nodes in T_h not on $\partial\Omega$.

The test space is chosen to be $V_h = \text{span}\{\xi_i\}_1^{N'}$ where ξ_i is piecewise constant given by

$$\xi_i = \begin{cases} 1 & \text{on } d_i, \\ 0 & \text{otherwise.} \end{cases}$$

To construct the trial space U_h, we follow the discussion in [12] based on the idea of exponential fitting proposed independently in [1] and [19]. For each $e_{i,j} \in E_h$ connecting the two neighboring nodes x_i and x_j, we define an exponential function $\phi_{i,j}$ on $e_{i,j}$ by

$$\begin{aligned} &\frac{d}{d\boldsymbol{e}_{i,j}}(\varepsilon\frac{d\phi_{i,j}}{d\boldsymbol{e}_{i,j}} - \bar{a}_{i,j}\phi_{i,j}) = 0, \\ &\phi_{i,j}(x_i) = 1, \quad \phi_{i,j}(x_j) = 0, \end{aligned} \tag{3.1}$$

where $\boldsymbol{e}_{i,j}$ denotes the unit vector from x_i to x_j and $\bar{a}_{i,j}$ is a constant approximation to $\boldsymbol{a}\cdot\boldsymbol{e}_{i,j}$ on $e_{i,j}$ such that the mapping $\boldsymbol{a}\cdot\boldsymbol{e}_{i,j}\mapsto\bar{a}_{i,j}$ from $C(e_{i,j}) \mapsto P_0(e_{i,j})$ preserves constants (e.g., $\bar{a}_{i,j} = (\boldsymbol{a}(x_i) + \boldsymbol{a}(x_j))\cdot\boldsymbol{e}_{i,j}/2$), where $C(e_{i,j})$ and $P_0(e_{i,j})$ denote respectively the spaces of all continuous functions and all 0th order polynomials on $e_{i,j}$. The above linear, constant coefficient two-point boundary-value problem can be solved exactly, yielding the local 1D basis function $\phi_{i,j}$ on the edge $e_{i,j}$. We then extend $\phi_{i,j}$ to $b_{i,j}$ by defining it to be constant along perpendiculars to $e_{i,j}$. Using this exponential function, we define a global basis function for U_h on Ω as follows:

$$\phi_i = \begin{cases} \phi_{i,j} & \text{on } b_{i,j} \text{ if } j \in I_i, \\ 0 & \text{otherwise,} \end{cases}$$

where $b_{i,j}$ denotes the element of B_h containing $e_{i,j}$ and

$$I_i = \{j : e_{i,j} \in E_h'\}$$

denotes the index set of all neighboring nodes of x_i. The support of ϕ_i is star-shaped. We put $U_h = \text{span}\{\phi_i\}_1^{N'}$. Obviously, we have $U_h \subset L^2(\Omega)$ and thus the trial space is nonconforming. This finite element space has the property that for any sufficiently smooth function u, the projection of the flux of the U_h-interpolant u_I of u on $\boldsymbol{e}_{i,j}$ satisfies

$$f_{i,j} := \varepsilon \frac{du_I}{d\boldsymbol{e}_{i,j}} - \bar{a}_{i,j} u_I = \frac{\varepsilon}{|e_{i,j}|} \left(B(\frac{\bar{a}_{i,j}|e_{i,j}|}{\varepsilon}) u_j - B(-\frac{\bar{a}_{i,j}|e_{i,j}|}{\varepsilon}) u_i \right) \tag{3.2}$$

on the edge $e_{i,j}$, where B denotes the Bernoulli function defined by

$$B(x) = \begin{cases} \dfrac{x}{e^x - 1}, & x \neq 0, \\ 1, & x = 0. \end{cases} \tag{3.3}$$

To define the discrete problem using U_h and V_h, it is convenient to introduce some notation. Let $l_{i,j} = \partial d_i \cap \partial d_j$ denote the intersection of the boundaries of d_i and d_j (cf., Fig. 3). Clearly, $|l_{i,j}| = \frac{2|b_{i,j}|}{|e_{i,j}|}$ if $j \in I_i$ and $|l_{i,j}| = 0$ otherwise. Corresponding to each $l_{i,j}$ we introduce a unit vector $\boldsymbol{l}_{i,j}$, which is directed so that $\arg(\boldsymbol{l}_{i,j}) = \arg(\boldsymbol{e}_{i,j}) + \pi/2$. For convenience we let $\hat{\boldsymbol{a}}$ be the approximation of $\boldsymbol{a}$ defined on $\bar{\Omega}$ such that, for all $b_{i,j} \in B_h$

$$\hat{\boldsymbol{a}}|_{b_{i,j}} = \bar{a}_{i,j} \boldsymbol{e}_{i,j} + \bar{a}_{i,j}^{\perp} \boldsymbol{l}_{i,j}, \tag{3.4}$$

where $\bar{a}_{i,j}$ is the constant used in (3.1) and $\bar{a}_{i,j}^{\perp}$ is a constant approximation to $\boldsymbol{a} \cdot \boldsymbol{l}_{i,j}$. Obviously, $\hat{\boldsymbol{a}}$ is a piecewise constant approximation to $\boldsymbol{a}$ on Ω. Note that the component $\bar{a}_{i,j}^{\perp}$ will make no contribution to the rest of the paper, but it allows us to use the convenient notation $\hat{\boldsymbol{a}}$ in some discussion. Before defining the finite element problem, we first introduce the mass lumping operator $P : C(\bar{\Omega}) \mapsto V_h$ such that

$$P(w)(x) = \sum_{i=1}^{N'} w(x_i) \xi_i(x), \qquad x \in \bar{\Omega}, \tag{3.5}$$

for all $w \in C(\bar{\Omega})$. It is easy to show that the mass-lumping mapping P is surjective from U_h to V_h because any function in U_h or V_h is determined uniquely by its nodal values at the vertices of T_h not on $\partial\Omega$. Using this mapping, we define the following finite element problem:

Problem 3.1. *Find $u_h \in U_h$ such that*

$$a(u_h, v_h) + (P(Gu_h), v_h) = (F, v_h) \quad \forall v_h \in V_h,$$

where $a(\cdot, \cdot)$ is a bilinear form defined by

$$a(u_h, v_h) = -\sum_{d \in D_h'} \int_{\partial d} (\varepsilon \nabla u_h - \hat{\boldsymbol{a}} u_h) \cdot \boldsymbol{n} v_h|_d ds. \tag{3.6}$$

Here $v_h|_d$ denotes the restriction of v_h on d.

Clearly, $a(\cdot,\cdot)$ is a nonstandard bilinear form. It is motivated by the resulting form of multiplying both sides of (2.1) by a piecewise constant function on Ω (i.e., an element in V_h) and then integrating by parts.

Because the mass lumping operator P defined in (3.5) is surjective from U_h to V_h, Problem 3.1 can be rewritten as the following Bubnov-Galerkin problem:

Problem 3.2. *Find $u_h \in U_h$ such that*

$$b(u_h, v_h) = (F, P(v_h)) \quad \forall v_h \in U_h,$$

where $b(\cdot,\cdot)$ is a bilinear form on $U_h \times U_h$ defined by

$$b(u_h, v_h) := a(u_h, P(v_h)) + (P(Gu_h), P(v_h)). \tag{3.7}$$

4. Convergence

In this section we consider the convergence of the approximate solution to Problem 3.2. We first show that the bilinear form defined in (3.7) is coercive. This implies that Problem 3.2 is uniquely solvable and the finite element formulation is numerically stable. We will then use the coercivity result to show that the approximation error of the finite element solution in a discrete norm is bounded above by $\mathcal{O}(h^{1/2})$.

Before further discussion, we first make the following assumption.

Assumption 4.1. *Assume that the mesh size h is sufficiently small and $\hat{\boldsymbol{a}}$ is properly chosen such that the inequality*

$$\frac{1}{2}\int_{\partial d_i} \hat{\boldsymbol{a}} \cdot \boldsymbol{n} ds + G(x_i)|d_i| \geqslant 0 \tag{4.1}$$

holds for all $d_i \in D_h'$, where $\hat{\boldsymbol{a}}$ is the approximation of $\boldsymbol{a}$ defined in (3.4) and x_i denotes the mesh node contained in d_i. Also, there exists a positive constant C, independent of h, such that for any $e_{i,j} \in E_h^{(1)}$,

$$\textit{either} \quad \boldsymbol{a}(x) \cdot \boldsymbol{e}_{i,j} = 0 \quad \textit{or} \quad |\boldsymbol{a}(x) \cdot \boldsymbol{e}_{i,j}| \geqslant C, \tag{4.2}$$

for all $x \in b_{i,j}$.

We comment that (4.1) is essentially a discrete analogue of (2.4). In fact, it can be obtained by integrating (2.4) over $d_i \in D_h'$, applying the integration by parts to the first term and then approximating $\boldsymbol{a}$ and G by $\hat{\boldsymbol{a}}$ and $G(x_i)$ respectively. Since the approximation $\bar{a}_{i,j} = \hat{\boldsymbol{a}} \cdot \boldsymbol{e}_{i,j}$ on any edge $e_{i,j}$ is rather arbitrary, it is possible to choose a_i properly so that (4.1) is satisfied when h is sufficiently small. The second condition (4.2) requires that the mesh is aligned in such a way that in each $b_{i,j}$, the edge $e_{i,j}$ is either perpendicular to the characteristic direction or the smaller angle between them is uniformly away from $\pi/2$. Clearly, (4.2) implies that for any $e_{i,j} \in E_h^{(1)}$

$$\text{either} \quad \bar{a}_{i,j} = 0 \quad \text{or} \quad |\bar{a}_{i,j}| \geqslant a_0 \tag{4.3}$$

for some $a_0 > 0$, independent of h.

Furthermore, since all the mesh lines in $T_h^{(2)}$ are parallel to one of the axes and $\boldsymbol{a}$ satisfies (2.5), it is obvious that

$$\min_{e_{i,j}\in E_h^{(2)}} |\hat{\boldsymbol{a}}\cdot\boldsymbol{e}_{i,j}| = \min_{e_{i,j}\in E_h^{(2)}} |\bar{a}_{i,j}| \geqslant \min\{\alpha_1,\alpha_2\}. \tag{4.4}$$

Let $\kappa_{i,j}$ be defined by

$$\kappa_{i,j} = \begin{cases} |e_{i,j}|, & |\bar{a}_{i,j}| \geqslant a_0, \\ \varepsilon, & |\bar{a}_{i,j}| = 0. \end{cases}$$

We now define two functionals, $||\cdot||_h$ and $||\cdot||$ respectively by

$$\|u_h\|_h^2 := \sum_{e_{i,j}\in E_h'} \kappa_{i,j}\left(\frac{u_j-u_i}{|e_{i,j}|}\right)^2 |b_{i,j}| \tag{4.5}$$

and

$$||u_h||^2 := ||u_h||_h^2 + \sum_{i=1}^{N'} u_i^2\left(\frac{1}{2}\int_{\partial d_i}\hat{\boldsymbol{a}}\cdot\boldsymbol{n}ds + G_i|d_i|\right)$$

for each $u_h = \sum_{i=1}^{N'} u_i\phi_i \in U_h$.

In the following lemma we show that both of these are norms on U_h.

Lemma 4.1. *The functional $||\cdot||_h$ is a norm on U_h. Furthermore, if Assumption 4.1 is fulfilled, then $||\cdot||$ is also a norm on U_h.*

Proof. The proof that (4.5) is a norm is easy and thus omitted here. The proof that $||\cdot||$ is a norm is just a consequence of the first part of this lemma and (4.1). □

The following theorem shows that the bilinear form $b(\cdot,\cdot)$ is coercive with respect to the norm $||\cdot||$.

Theorem 4.1. *Let Assumption 4.1 be fulfilled. Then, for all $u\in U_h$, we have*

$$b(u,u) \geqslant C||u||^2, \tag{4.6}$$

where C denotes a positive constant independent of ε, h, and u.

Proof. For any $u\in U_h$ it is shown in Section 4 of [12] that

$$\begin{aligned} a(u_h,P(u_h)) &= -\sum_{d\in D_h'}\int_{\partial d}(\varepsilon\nabla u_h - \hat{\boldsymbol{a}}u_h)\cdot\boldsymbol{n}P(u_h)ds \\ &= \sum_{e_{i,j}\in E_h'}\frac{\sigma_{i,j}}{2}\frac{\rho_{i,j}}{e^{\rho_{i,j}}-1}(1+e^{\rho_{i,j}})(u_j-u_i)^2\frac{2|b_{i,j}|}{|e_{i,j}|} \\ &\quad + \sum_{e_{i,j}\in E_h'}\frac{\bar{a}_{i,j}}{2}(u_i^2-u_j^2)|l_{i,j}| \end{aligned} \tag{4.7}$$

where $l_{i,j} = \partial d_i\cap\partial d_j$, $\sigma_{i,j} = \varepsilon/|e_{i,j}|$, $\rho_{i,j} = \bar{a}_{i,j}/\sigma_{i,j}$ and $B(\cdot)$ is the Bernoulli function defined in (3.3). We consider the following two cases.

Case 1: $|\bar{a}_{i,j}| \geqslant a_0$.

Since

$$\frac{(e^{\rho_{i,j}}+1)}{(e^{\rho_{i,j}}-1)}\bar{a}_{i,j} \geqslant |\bar{a}_{i,j}|.$$

We have

$$\begin{aligned}\sum_{e_{i,j}\in E_h'} \frac{\sigma_{i,j}}{2}\frac{\rho_{i,j}}{e^{\rho_{i,j}}-1}(1+e^{\rho_{i,j}})(u_j-u_i)^2\frac{2|b_{i,j}|}{|e_{i,j}|} &\geqslant C\sum_{e_{i,j}\in E_h'}|e_{i,j}||\bar{a}_{i,j}|\left(\frac{u_j-u_i}{|e_{i,j}|}\right)^2|b_{i,j}| \\ &\geqslant C\sum_{e_{i,j}\in E_h'}|e_{i,j}|\left(\frac{u_j-u_i}{|e_{i,j}|}\right)^2|b_{i,j}|,\end{aligned}$$

since $|\bar{a}_{i,j}| \geqslant a_0$.

Case 2: $|\bar{a}_{i,j}| = 0$.

When $|\bar{a}_{i,j}| = 0$, $\rho_{i,j} = 0$ and

$$\lim_{\rho_{i,j}\to 0}\frac{\rho_{i,j}}{e^{\rho_{i,j}}-1} = B(0) = 1$$

by (3.3). Therefore,

$$\sum_{e_{i,j}\in E_h'} \frac{\sigma_{i,j}}{2}\frac{\rho_{i,j}}{e^{\rho_{i,j}}-1}(1+e^{\rho_{i,j}})(u_j-u_i)^2\frac{2|b_{i,j}|}{|e_{i,j}|} \geqslant C\sum_{e_{i,j}\in E_h'}\varepsilon\left(\frac{u_j-u_i}{|e_{i,j}|}\right)^2|b_{i,j}|.$$

Combining the above two cases, we have

$$\begin{aligned}\sum_{e_{i,j}\in E_h'} \frac{\sigma_{i,j}}{2}\frac{\rho_{i,j}}{e^{\rho_{i,j}}-1}(1+e^{\rho_{i,j}})(u_j-u_i)^2\frac{2|b_{i,j}|}{|e_{i,j}|} &\geqslant C\sum_{e_{i,j}\in E_h'}\kappa_{i,j}\left(\frac{u_j-u_i}{|e_{i,j}|}\right)^2|b_{i,j}| \\ &= C||u_h||_h^2.\end{aligned} \tag{4.8}$$

Now, let us consider the last term in (4.7). Transforming from a summation over the edges to a summation over the nodes of X_h',

$$\sum_{e_{i,j}\in E_h'}\frac{\bar{a}_{i,j}}{2}(u_i^2-u_j^2)|l_{i,j}| = \frac{1}{2}\sum_{i=1}^{N'}u_i^2\left(\sum_{j\in I_i}\bar{a}_{i,j}|l_{i,j}|\right) = \frac{1}{2}\sum_{i=1}^{N'}u_i^2\int_{\partial d_i}\hat{\boldsymbol{a}}\cdot\boldsymbol{n}ds\,.$$

Therefore, substituting the above inequality and (4.8) into (4.7), we obtain from (3.7)

$$\begin{aligned}b(u_h,u_h) &= a(u_h,P(u_h)) + (P(Gu_h),P(u_h)) \\ &\geqslant C||u_h||_h^2 + \sum_{i=1}^{N'}u_i^2\left(\frac{1}{2}\int_{\partial d_i}\hat{\boldsymbol{a}}\cdot\boldsymbol{n}ds + G_i|d_i|\right) \\ &\geqslant C||u_h||^2.\end{aligned}$$

□

We now establish an upper bound for $||u_I - u_h||$, where u_I and u_h denote respectively the U_h-interpolant of the solution u to Problem 2.1 and the solution to Problem 3.1. We start this discussion by the following lemma that will be used in the proof of the main result of this section.

Lemma 4.2. *Let Assumption 4.1 be fulfilled. Then, for any sufficiently smooth function u, the approximate flux defined in (3.2) satisfies*

$$||\boldsymbol{f}\cdot\boldsymbol{e}_{i,j}-f_{i,j}||_{\infty,b_{i,j}}\leqslant\begin{cases}C|e_{i,j}|\big(|\boldsymbol{f}|_{1,\infty,b_{i,j}}+|\boldsymbol{a}|_{1,\infty,b_{i,j}}||u||_{\infty,b_{i,j}}\big), & |\boldsymbol{a}\cdot\boldsymbol{e}_{i,j}|\geqslant a_0 \text{ on } b_{i,j},\\ C\varepsilon|e_{i,j}||u|_{2,\infty,b_{i,j}}, & \boldsymbol{a}\cdot\boldsymbol{e}_{i,j}=0 \text{ on } b_{i,j},\end{cases}\tag{4.9}$$

where C is a positive constant, independent of h,ε, and u.

Proof. Let C be a generic constant, independent of h, ε, and u. When $|\boldsymbol{a}\cdot\boldsymbol{e}_{i,j}|\geqslant C$, $|\bar{a}_{i,j}|\geqslant a_0$ by (4.3). Then on $e_{i,j}$,

$$\boldsymbol{f}\cdot\boldsymbol{e}_{i,j}-f_{i,j}=\left(\left(\varepsilon\frac{du}{d\boldsymbol{e}_{i,j}}-\bar{a}_{i,j}u\right)-f_{i,j}\right)+(\bar{a}_{i,j}-\boldsymbol{a}\cdot\boldsymbol{e}_{i,j})u\,.$$

Because the mappings $(\varepsilon\frac{du}{d\boldsymbol{e}_{i,j}}-\bar{a}_{i,j}u)\mapsto f_{i,j}$ and $\boldsymbol{a}\cdot\boldsymbol{e}_{i,j}\mapsto\bar{a}_{i,j}$ preserve constants, we have (cf., [4], Theorem 3.1.4) from the above equality

$$||\boldsymbol{f}\cdot\boldsymbol{e}_{i,j}-f_{i,j}||_{\infty,e_{i,j}}\leqslant C|e_{i,j}|\big(|\boldsymbol{f}|_{1,\infty,e_{i,j}}+|\boldsymbol{a}|_{1,\infty,e_{i,j}}||u||_{\infty,e_{i,j}}\big).$$

This inequality can then be extended to the element $b_{i,j}$, yielding the first case in (4.9).

When $\boldsymbol{a}\cdot\boldsymbol{e}_{i,j}=0$ on $b_{i,j}$, we have $\bar{a}_{i,j}=0$, and so

$$f_{i,j}=\frac{\varepsilon}{|e_{i,j}|}(u_j-u_i),$$

which is a difference approximation to $\varepsilon\nabla u\cdot\boldsymbol{e}_{i,j}$ on $e_{i,j}$. Therefore,

$$||\boldsymbol{f}\cdot\boldsymbol{e}_{i,j}-f_{i,j}||_{\infty,b_{i,j}}=\left\|\varepsilon\nabla u\cdot\boldsymbol{e}_{i,j}-\varepsilon\frac{u_j-u_i}{|e_{i,j}|}\right\|_{\infty,b_{i,j}}\leqslant C\varepsilon|e_{i,j}||u|_{2,\infty,b_{i,j}}.$$

This completes the proof. □

We now state the following lemma without proving.

Lemma 4.3. *Let Assumption 2.1 be fulfilled. If $\beta\geqslant 2$ in (2.6) and $M_1=M_2=M$, a positive integer, then, for any element edge $e_{i,j}\in E_h^{(2)}$, the set of edges in $T_h^{(2)}$, there exists a positive constant C, independent of h, u, and ε, such that*

$$\int_{l_{i,j}}|\boldsymbol{f}\cdot\boldsymbol{e}_{i,j}-f_{i,j}|ds\leqslant\begin{cases}C|l_{i,j}|hK_1, & e_{i,j}\subset\bar{\Omega}_2^{(1)}\cup\bar{\Omega}_2^{(2)},\\ C|l_{i,j}|\dfrac{1}{M}\ln\dfrac{1}{\varepsilon}, & e_{i,j}\subset\Omega_2^{(3)},\end{cases}$$

where $\boldsymbol{f}$ and $f_{i,j}$ are defined in (2.2) and (3.2), respectively, and

$$K_1=\max\left\{1,h^{-1}\varepsilon^{\beta/2M},h^{-1}\varepsilon\ln\frac{1}{\varepsilon}\right\}.\tag{4.10}$$

Proof. The proof of this lemma which uses (4.4) can be found in [24]. □

The upper error bound for the error in the finite element solution is established the following theorem.

Theorem 4.2. *Let Assumptions 2.1 and 4.1 be fulfilled. If $\beta \geqslant 2$ in (2.6) and $M_1 = M_2 = M$, a positive integer, then there exists a positive constant C, independent of h, u, and ε, such that*

$$||u_I - u_h|| \leqslant Ch^{1/2}\left(1 + h^{1/2}K_2\right), \tag{4.11}$$

where u_I and u_h denote respectively the U_h-interpolant of the solution u to Problem 2.1 and the solution to Problem 3.1, and K_2 is defined as

$$K_2 = \max\left\{M^{1/2}K_1, M^{-1/2}\sqrt{\varepsilon}\ln\frac{1}{\varepsilon}\right\}, \tag{4.12}$$

with K_1 being defined in (4.10).

Proof. Let C denote a generic positive constant, independent of h, ε, and u. For any $v_h \in U_h$, multiplying the continuous equation (2.1) by $P(v_h)$ and integrating the first term by parts, we have

$$-\sum_{d\in D'_h}\int_{\partial d} \boldsymbol{f}\cdot\boldsymbol{n}P(v_h)ds + (Gu, P(v_h)) = (F, P(v_h)).$$

From this and Problem 3.2 we obtain

$$\begin{aligned} a(u_h, P(v_h)) + (P(Gu_h), P(v_h)) &= (F, P(v_h)) \\ &= -\sum_{d\in D'_h}\int_{\partial d} \boldsymbol{f}\cdot\boldsymbol{n}P(v_h)ds + (Gu, P(v_h)). \end{aligned}$$

Taking $a(u_I, P(v_h)) + (P(Gu_I), P(v_h))$ away from both sides of this equation gives

$$\begin{aligned} &a(u_h - u_I, P(v_h)) + (P(Gu_h) - P(Gu_I), P(v_h)) \\ &\qquad = -\sum_{d\in D'_h}\int_{\partial d} \boldsymbol{f}\cdot\boldsymbol{n}P(v_h)ds - a(u_I, P(v_h)) + (Gu - P(Gu_I), P(v_h)). \end{aligned}$$

So, using (3.7) and taking absolute value on both sides of the above, we have

$$\begin{aligned} |b(u_h - u_I, v_h)| &\leqslant \left|-\sum_{d\in D'_h}\int_{\partial d} \boldsymbol{f}\cdot\boldsymbol{n}P(v_h)ds - a(u_I, P(v_h))\right| + |(Gu - P(Gu_I), P(v_h))| \\ &=: R_1 + R_2. \end{aligned} \tag{4.13}$$

We now consider the two error terms R_1 and R_2 separately.

For R_1, using (3.6) and transforming from a summation over the nodes to one over the edges, we have

$$R_1 = \left|-\sum_{d\in D'_h}\int_{\partial d} (\boldsymbol{f} - (\varepsilon\nabla u_I - \hat{\boldsymbol{a}}u_I))\cdot\boldsymbol{n}P(v_h)ds\right| \leqslant \left|\sum_{e_{i,j}\in E'_h}(v_i - v_j)\int_{l_{i,j}}(\boldsymbol{f}\cdot\boldsymbol{e}_{i,j} - f_{i,j})ds\right|.$$

We split E_h' into two parts: $E_h^{(1)}$ and $E_h^{(2)}$ corresponding to $T_h^{(1)}$ and $T_h^{(2)}$, respectively, and so $R_1 =: R_1^{(1)} + R_1^{(2)}$. Using the Cauchy–Schwarz inequality, Theorem 2.1, the error bounds (4.9) in Lemma 4.2 and $|l_{i,j}| = 2|b_{i,j}|/|e_{i,j}|$, we have

$$\begin{aligned}
R_1^{(1)} &\leqslant \sum_{e_{i,j}\in E_h^{(1)}} |v_j - v_i| \int_{l_{i,j}} |(\boldsymbol{f}\cdot\boldsymbol{e}_{i,j} - f_{i,j})| ds \\
&\leqslant C \sum_{e_{i,j}\in E_h^{(1)}} |v_j - v_i| \sup_{x\in b_{i,j}} |(\boldsymbol{f}\cdot\boldsymbol{e}_{i,j} - f_{i,j})| \frac{2|b_{i,j}|}{|e_{i,j}|} \\
&\leqslant C \sum_{e_{i,j}\in E_h^{(1)}} \kappa_{i,j} \frac{|v_j - v_i|}{|e_{i,j}|} |b_{i,j}| \qquad (4.14) \\
&\leqslant C \left(\sum_{e_{i,j}\in E_h^{(1)}} \kappa_{i,j} \left(\frac{v_j - v_i}{|e_{i,j}|} \right)^2 |b_{i,j}| \right)^{1/2} \left(\sum_{e_{i,j}\in E_h^{(1)}} \kappa_{i,j} |b_{i,j}| \right)^{1/2} \\
&\leqslant C h^{1/2} ||v_h||_h,
\end{aligned}$$

since $\kappa_{i,j} \leqslant h$.

For the term $R_1^{(2)}$, using Lemma 4.3, it can be shown (cf., [24]) that

$$R_1^{(2)} \leqslant ChK_2 \left[\sum_{e_{i,j}\in E_h^{(2)}} |e_{i,j}| \left(\frac{v_j - v_i}{e_{i,j}} \right)^2 |b_{i,j}| \right]^{1/2} \leqslant CK_2 ||v_h||_h \,. \qquad (4.15)$$

Let us now consider R_2 in (4.13). This term can also be split into two terms as follows:

$$R_2 = (Gu - P(Gu_I), P(v_h))_{\Omega_1} + (Gu - P(Gu_I), P(v_h))_{\Omega_2} := R_2^{(1)} + R_2^{(2)}.$$

Since the mapping P preserves constants, using the standard finite element interpolation argument [17, Theorem 6.8], we have

$$R_2^{(1)} \leqslant Ch|Gu|_1 ||v_h||_\infty \leqslant Ch||v_h||_\infty \qquad (4.16)$$

because of Theorem 2.1. It has been shown in [24] that

$$R_2^{(2)} \leqslant Ch||v_h||_\infty \left(\varepsilon \ln \frac{1}{\varepsilon} + e^{\beta/2M} \right) \leqslant Ch||v_h||_\infty.$$

Substituting (4.14), (4.15), (4.16) and the above inequality into (4.13), we have

$$|b(u_h - u_I, v_h)| \leqslant C \left[h^{1/2}(1 + h^{1/2}K_2)||v_h|| + h||v_h||_\infty \right].$$

Choosing $v_h = u_h - u_I$ and using (4.6), we obtain

$$||u_h - u_I||^2 \leqslant C \left[h^{1/2}(1 + h^{1/2}K_2)||u_h - u_I|| + h \right],$$

since $||u_h - u_I||_\infty \leqslant C$. This is of the form

$$y^2 \leqslant C(1 + h^{1/2}K_2)h^{1/2}y + Ch$$

or

$$\left(y - \frac{1}{2}C(1+h^{1/2}K_2)h^{1/2}\right)^2 \leqslant Ch + \frac{(C(1+h^{1/2}K_2))^2}{4}h.$$

The above reduces to

$$y \leqslant \sqrt{Ch + \frac{(C(1+h^{1/2}K_2))^2}{4}h} + \frac{1}{2}C(1+h^{1/2}K_2)h^{1/2} \leqslant Ch^{1/2}\left(1+h^{1/2}K_2\right).$$

Replacing y with $||u_h - u_I||$, we obtain

$$||u_h - u_I|| \leqslant Ch^{1/2}\left(1+h^{1/2}K_2\right).$$

This completes the proof of the theorem. □

Corollary 4.1. *Let the assumptions in Theorem 4.2 be fulfilled and assume that $\varepsilon \ll h$. Then, there exists a positive constant C, independent of h, ε, and u, such that the following results hold.*

1. *If M is chosen such that $h^{-1}\varepsilon^{\beta/2M} \leqslant \mathcal{O}(1)$, then we have*

$$||u_I - u_h|| \leqslant Ch^{1/2}(1+h^{1/2}M^{1/2}). \tag{4.17}$$

2. *If M is such that $h^{-1}\varepsilon^{\beta/2M} = \mathcal{O}(1)$, then*

$$||u_I - u_h|| \leqslant Ch^{1/2}\left(1+h^{1/2}\sqrt{\left|\frac{\ln\varepsilon}{\ln h}\right|}\right). \tag{4.18}$$

3. *If $M = 1$, then,*

$$||u_I - u_h|| \leqslant Ch^{1/2}. \tag{4.19}$$

Proof. Let C be a generic constant, independent of h, u, and ε. When $\varepsilon \ll h$, $h^{-1}e\ln(1/\varepsilon) < C$. From the definition of K_1 and K_2 in (4.10) and (4.12), respectively, we see that in this case, $K_2 \leqslant C$. Therefore, (4.17) follows from (4.11) and $K_2 \leqslant C$.

To prove item 2, we note that $h^{-1}\varepsilon^{\beta/2M} = 1$ implies $M = \left\lceil \frac{\beta\ln\varepsilon}{2\ln h} \right\rceil$. Thus, (4.17) implies (4.18).

Finally, if $M = 1$ and $\varepsilon \ll h < 1$, then

$$h^{-1}\varepsilon^{\beta/2M} \leqslant h^{-1}\varepsilon < h^{-1}\varepsilon\ln(1/\varepsilon) < C.$$

So, (4.19) follows from these inequalities and (4.11). □

Remark 4.1. We comment that it would be thought that the error bound in (4.17) is independent of ε. But in fact, M is dependent on ε when h and ε are given. However, if a Shishkin-type mesh is used, i.e., the transition widths in (2.6) are chosen according to the Shishkin law [16], then the resulting error bounds may be independent of ε. We will investigate this in the near future.

Remark 4.2. Though the right-hand side of (4.18) is still a function of ε, it depends very weakly on ε. This can be seen by considering the equation $h^{1/2}\sqrt{\ln\varepsilon/\ln h} = 1$ when $h = 0.1$. It has the solution

$$\varepsilon = e^{-10\times\ln 10} \approx 10^{-11}.$$

Therefore, even for this very coarse mesh, $h^{1/2}\sqrt{\ln\varepsilon/\ln h}$ remains bounded for the values of ε as small as 10^{-11}. This implies that the error bound in (4.18) is independent of ε for any practical range of the values of ε.

Remark 4.3. We also remark that (4.19) offers an ε-uniform convergence at the rate $\mathcal{O}(h^{1/2})$. In this case, all the interior mesh points are away from the layers and thus the box method does not resolve the layers.

Remark 4.4. We comment that notable advances on the uniform convergence of the streamline-diffusion method and the standard piecewise bilinear finite element method are given in [20, 21] and [25], respectively. All of these existing works are based on structured Shishkin type piecewise uniform rectangular partitions. The error bounds obtained in both [21] and [25] are of order $\bar{N}^2\ln\bar{N}$ in the respective energy norms, where $\bar{N}$ denotes the number of mesh points along the x and y directions. Clearly, these error bounds are better than what we have established in this paper due to the supercovergence. While [21] and [25] focus on the superconvergence phenomena of the methods on piecewise uniform meshes of Shishkin type, the aim of this paper is, however, to provide realistic error estimates for the finite volume method on unstructured meshes of a combination of triangles and rectangles. The error estimates and the stability have been established without using any unrealistic assumptions on the meshes such as that of 'no-obtuse angles'. However, if a mesh is favorably aligned, the finite volume method may provide a convergence rate higher than $\mathcal{O}(h^{1/2})$, as demonstrated numerically in the next section. Furthermore, the (discrete) energy norm used here is virtually ε-independent, while energy norms in most of the existing work depend on ε. Therefore, the stability of the method is independent of ε. Unlike the streamline-diffusion method discussed in [20,21], the finite volume method does not depend on any users' chosen parameter. Also, the computed fluxes are locally conservative because the finite volume method is based on the local conservation law. This is important for solving problems such as semiconductor device equations (cf., [3, 23]).

To conclude this section, we note that the finite volume method has been applied to many real-world problems, we will not compare it numerically with some other existing ones. Instead, we only present, in the next section, some numerical results to verify the theoretical rates of convergence obtained in the previous sections.

5. Numerical experiments

To demonstrate the theoretical results obtained in the previous sections, numerical experiments on the following test problem have been performed. All the computations were carried out in double precision on a Pentium PC under the Cygwin environment.

Test. The test problem is chosen to be the following PDEs:

$$\begin{aligned} -\nabla\cdot(\varepsilon\nabla u - \boldsymbol{a}u) &= F \quad \text{in } \Omega = (0,1)^2, \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

with the exact solution being given by

$$u_{\text{exact}} = x^2y^2\left(1-\exp\left(\frac{x-1}{\varepsilon}\right)\right)\left(1-\exp\left(\frac{y-1}{\varepsilon}\right)\right)$$

and $\boldsymbol{a} = (1,1)^\top$.

This problem has two exponential layers at $x = 1$ and $y = 1$. Let us first look at the computed rates of convergence of the method for mesh points away from the layers by using conventional quasiuniform meshes. This corresponds to the case of $M = 1$ in Corollary 4.1. To do so, we choose a sequence of 6 Delaunay triangulations starting with an initial mesh with 26 nodes. This initial mesh contains 9 interior nodes and 17 boundary nodes, and the maximum and minimum angles in the mesh are about 147.7^0 and 15.8^0 respectively. This mesh is then refined five times by dividing a triangle in the mesh into four subtriangles by connecting the midpoints of the three edges of the triangle. We denote this sequence by T_{h_k} for $k = 1, 2, \ldots, 6$ with $h_1 \approx 1/4$. The approximated rate of convergence is defined as follows. For $k = 1, 2, \ldots, 5$, we define

$$p_k = \log_2 \frac{||u_{h_k} - u_{\text{exact}}||}{||u_{h_{k+1}} - u_{\text{exact}}||}.$$

We then define the computed rate of convergence to be $p = \sum_{i=1}^{5} p_k/5$, i.e., the average of the five approximations. Based on this definition, we also define the computed rates of convergence in the discrete maximum norm

$$||u_{h_k} - u_{\text{exact}}||_\infty = \max_{1\leqslant i\leqslant N} |u_i - u_{\text{exact}}(x_i)|$$

and the discrete L^2 norm

$$||u_{h_k} - u_{\text{exact}}||_0^2 = \sum_{i=1}^{N} (u_i - u_{\text{exact}}(x_i))^2 |d_i|.$$

The computed values of p are listed in Table 1. From the table it is seen that the rates of convergence in the discrete energy norm $||\cdot||$ are about 0.55 when ε is small.

Table 1. Computed rates of convergence in the three norms

ε	$\|\|\cdot\|\|$	$\|\|\cdot\|\|_\infty$	$\|\|\cdot\|\|_0$
1	1.98	1.92	2.01
10^{-1}	1.79	1.76	1.84
10^{-2}	1.07	0.84	1.23
10^{-3}	0.55	0.25	0.85
10^{-4}	0.57	0.26	0.82
10^{-5}	0.57	0.26	0.82
10^{-6}	0.57	0.26	0.82

To visualize the rates in a different way, we plot the computed errors in $||\cdot||$ for three different values of ε in Fig. 4, and compare them with the reference rate $h^{1/2}$. From the

figure, it is seen that the computed rates are close to $h^{1/2}$ when ε is small, as predicted in the previous section. When $\varepsilon = 0.01$, the computed rate is higher than $h^{1/2}$.

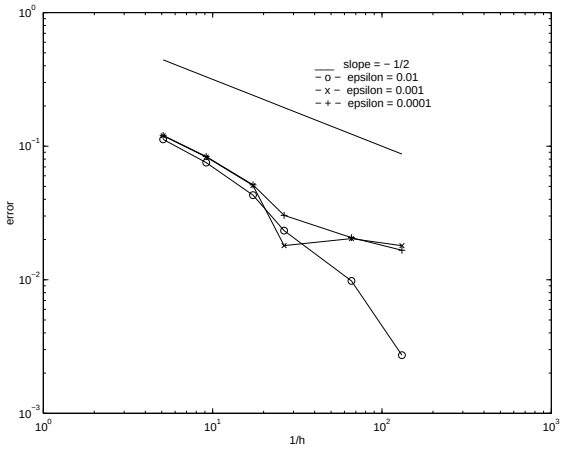

Figure 4. Computed errors in $||\cdot||$ on unstructured triangular meshes

We now consider the case of $M > 1$. In this investigation, we choose the transition parameter $\beta = 2$ in (2.6). A sequence of six uniform meshes for Ω_1 is chosen as follows. For a given positive integer, I_k, we first divide Ω_1 into a uniform rectangular mesh with $I_k \times I_k$ mesh points so that the mesh parameter is $h_k = 1/(I_k - 1)$. Each rectangle in the mesh is then divided into triangles by choosing one of the two diagonals. The mesh sequence corresponds to the sequence $\{h_k\}_1^6$ satisfying $h_1 = 1/4$ and $h_{k+1} = h_k/2$ for $k = 2, 3, 4, 5, 6$. For the subdomain Ω_2, we choose

$$M = \max\left\{\left\lceil\sqrt{I_k - 1}\right\rceil, \left\lceil\frac{\beta|\ln\varepsilon|}{2|\ln h|}\right\rceil, 1\right\}.$$

The computed rates of convergence in the three norms are listed in Table 2. Clearly, the numbers in the table show the phenomenon of superconvergence. In particular, the rates of convergence in $||\cdot||$ and $||\cdot||_\infty$ are about half an order higher than those in Table 1. This phenomenon may be because of the piecewise uniform partitions, though this superconvergence has not been proved mathematically. Certainly, it is worth further investigation.

Table 2. Computed rates of convergence using the piecewise uniform meshes

ε	$\|\|\cdot\|\|$	$\|\|\cdot\|\|_\infty$	$\|\|\cdot\|\|_0$
1	2.49	2.00	2.00
10^{-1}	1.46	1.07	1.31
10^{-2}	1.78	1.57	1.54
10^{-3}	1.05	0.90	1.04
10^{-4}	0.97	0.82	0.96
10^{-5}	0.96	0.81	0.95
10^{-6}	0.96	0.81	0.95

As in the previous case, we plot the computed errors in $||\cdot||$ for three different values of ε in Fig. 5, and compare them with the reference rate $\mathcal{O}(h)$. Clearly, from the figure we see

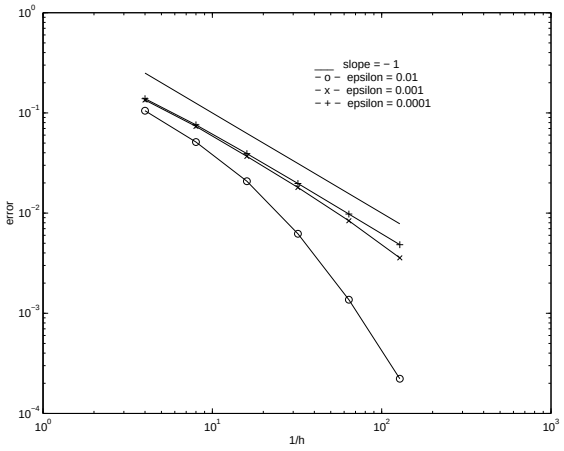

Figure 5. Computed errors in $||\cdot||$ on piecewise uniform meshes

that the computed rates are equal to $\mathcal{O}(h)$ when ε is small, and are higher than $\mathcal{O}(h)$ when $\varepsilon = 0.01$. These show the superconvergence phenomena due to the special alignments of the meshes.

6. Conclusions

In this paper we presented an analysis of the well-known exponentially fitted finite volume (or Scharfetter-Gummel box) method for a two-dimensional linear singularly perturbed convection-diffusion problem containing two exponential boundary layers. The method, constructed on a Delaunay partition containing triangles and rectangles, was first formulated as a Petrov-Galerkin finite element method, and then as a Bubnov-Galerkin finite element method. The stability and an $\mathcal{O}(h^{1/2})$ error estimate were established. It has been shown that the upper error bound is almost independent of the singularly perturbation parameter ε in general and independent of ε if all mesh internal nodes are away from the layers. Numerical results were presented to verify the theoretical rates of convergence. The numerical results suggested that, when piecewise uniform meshes are used, the method displays the behavior of superconvergence.

Acknowledgments

Part of the work was carried out while the first author was visiting Technische Universität Clausthal.

References

[1] D. N. de G. Allen and R. V. Southwell, *Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder*, Quart. J. Mech. Appl. Math., **8** (1955), pp. 129–145.

[2] L. Angermann, *Error estimates for the finite-element solution of an elliptic singularly perturbed problem*, IMA J. Num. Anal., **15** (1995), pp. 161–196.

[3] L. Angermann and S. Wang, *Three-dimensional exponentially fitted conforming tetrahedral finite elements for the semiconductor continuity equations*, Appl. Numer. Math., **46** (2003), pp. 19–43.

[4] P. G. Ciarlet, *The finite element method for elliptic problems*, North-Holland, Amsterdam, 1978.

[5] B. Delaunay, *Sur la sphère vide*, Izv. Akad. Nauk. SSSR., Math. and Nat. Sci. Div., **6** (1934), pp. 793–800.

[6] G. L. Dirichlet, *Über die Reduction der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen*, J. Reine Angew. Math., **40** (1850), No. 3, pp. 209–227.

[7] C. J. Fitzsimons, J. J. H. Miller, S. Wang, and C. H. Wu, *Hexahedral finite elements for the stationary semiconductor device equations*, Comp. Meth. Appl. Mech. Eng., **84** (1990), No. 5, pp. 43–57.

[8] T. Linß and M. Stynes, *Asymptotic analysis and Shishkin-type decomposition for an elliptic convection-diffusion problem*, J. Math. Anal. Appl., **261** (2001), pp. 604–632.

[9] B. J. McCartin, *Discretization of the Semiconductor Device Equations*, from *New Problems and New Solutions for Device and Process Modelling*, ed. J.J.H. Miller, Boole Press, Dublin, 1985.

[10] J. J. H. Miller and S. Wang, *A triangular mixed finite element method for the stationary semiconductor device equations*, RAIRO Modél. Math. Anal. Numér., **25** (1991), No. 4, pp. 441–463.

[11] J. J. H. Miller and S. Wang, *An ϵ-uniformly convergent finite element method for a singularly perturbed advection-diffusion equation*, in *Contributions in Numerical Mathematics*, ed. Agarwal R.P., World Scientific, Singapore, (1993) pp. 271–284.

[12] J. J. H. Miller and S. Wang, *A new non-conforming Petrov-Galerkin finite element method with triangular element for a singularly perturbed advection-diffusion problem*, IMA J. Numer. Anal., **14** (1994) pp. 257–276.

[13] J. J. H. Miller and S. Wang, *An exponentially fitted finite element volume method for the numerical solution of 2D unsteady incompressible flow problems*, J. Comput. Phys., **115** (1994), No. 1, pp. 56–64.

[14] J. J. H. Miller and S. Wang, *An Analysis of the Scharfetter-Gummel Box Method for the Stationary Semiconductor Device Equations*, RAIRO Modél. Math. Anal. Numér., **28** (1994), No. 2, pp. 123–140.

[15] M. S. Mock, *Analysis of a discretization algorithm for stationary continuity equations in semiconductor device models*, COMPEL, **2** (1983), pp. 117–139.

[16] J. J. H. Miller, E. O'Riordan, and G. I. Shishkin, *Fitted Numerical Methods for Singular Perturbation Problems*, World Scientific, Singapore, 1996.

[17] J. T. Oden and J. N. Reddy, *An Introduction to the Mathematical Theory of Finite Elements*, John Wiley & Sons, New York, 1976.

[18] H.-G. Roos, D. Adam, and A. Felgenhauer, *A novel non-conforming uniformly convergent finite element method in two dimensions*, J. Math. Anal. Appl., **201** (1996), pp. 711–755.

[19] D. Scharfetter and H. K. Gummel, *Large-signal analysis of a silicon read diode oscillator*, IEEE Trans. Elec. Dev., **ED-16** (1969), pp. 64–77.

[20] M. Stynes and L. Tobiska, *Analysis of the streamline-diffusion finite element method on a piecewise uniform mesh for a convection-diffusion problem with exponential layers*, East-West J. Numer. Math., **9** (2001), No. 1, pp. 59–76.

[21] M. Stynes, L. Tobiska, *The SDFEM for a convection-diffusion problem with a boundary layer: optimal error analysis and enhancement of accuracy*, SIAM J. Numer. Anal., (accepted).

[22] S. Wang, *A novel exponentially fitted triangular finite element method for an advection-diffusion problem with boundary layers*, J. Comp. Phys., **134** (1997), pp. 253–260.

[23] S. Wang, *A new exponentially fitted triangular finite element method for the continuity equations in the drift-diffusion model of semiconductor devices*, RAIRO Modél. Math. Anal. Numér., **33** (1999), No. 1, pp. 99–112.

[24] S. Wang and Z. C. Li, *A non-conforming combination of the finite element and volume methods with an anisotropic mesh refinement for a singularly perturbed convection-diffusion equation*, Math. Comp., **72** (2003), No. 244, pp. 1689–1709.

[25] Z. Zhang, *Finite element superconvergence on Shishkin mesh for 2D convection-diffusion problems*, Math. Comp., **72** (2003), No. 243, pp. 1147–1177.

Received 30 Nov. 2002
Revised 16 Jul. 2003